**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 17 June 2025 |
| Team ID | **AS PS VS VV** |
| Project Title | Unemployed Insurance Beneficiary Forecasting |
| Maximum Marks | 2 Marks |

**Data Quality Report Template**

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Source** | **Data Quality Issue** | **Severity** | **Resolution Plan** |
| Dataset | Presence of null values in some columns | Moderate | Used df.isnull().sum() to identify missing values. Applied df.dropna(inplace=True) to remove rows with nulls, ensuring clean data for modeling. |
| Dataset | Potential duplicate records | Low | Checked with df.duplicated().sum(). No duplicates found, so no further action required. |
| Dataset | Inconsistent region/county names | Low | Used df['region'].value\_counts() and standardized names if inconsistencies were found. |
| Dataset | Data type inconsistencies | Low | Verified with df.info(). Converted columns to correct data types as needed (e.g., dates to datetime). |